Overview of BioCreative III Gene Normalization

Zhiyong Lu[1], W. John Wilbur[1]

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD 20894 USA

Email addresses:

ZL: luzh@ncbi.nlm.nih.gov

WJW: wilbur@ncbi.nlm.nih.gov

Abstract

Background

The Gene Normalization (GN) task refers to the identification and linking of gene mentions in free text to standard gene database identifiers, an important task motivated by many real-world uses such as assisting literature curation for model organism databases. Here we report the GN challenge in BioCreative III where participating teams are asked to return a ranked list of gene ids of full-text articles. For training, we prepared 32 fully annotated articles and 500 partially annotated articles. A total of 507 articles were selected as the test set. We developed an EM algorithm approach for selecting 50 articles from the test set for obtaining gold-standard human annotations and used the same algorithm for inferring ground truth over the whole set of 507 articles based on team submissions. We report team performance by a newly proposed metric for measuring retrieval efficacy called Threshold Average Precision (TAP-k).

Results

We received a total of 37 runs from 14 different teams for the BioCreative III GN task. When evaluated using the gold-standard annotations of the 50 articles, the highest TAP-k scores are 0.3248 (k=5), 0.3469 (k=10), and 0.3466 (k=20), respectively. Higher TAP-k scores of 0.4581 (k=5, 10) and 0.4684 (k=20) are observed when evaluated using the inferred ground truth over the full test set.

Conclusions

Overall team results show that this year's GN task is more challenging than past events, which is likely due to the complexity of full text as well as species identification. By comparing team rankings with different evaluation data (gold

standard vs. inferred ground truth), we demonstrate that our approach succeeds in inferring ground truth adequate for effectively detecting good team performance.

Background

The gene normalization (GN) task in BioCreative III is similar to past GN tasks in BioCreative I and II (1-3) in that the goal is to link genes or gene products mentioned in the literature to standard database identifiers. This task has been inspired partly by a pressing need to assist model organism database (MOD) literature curation efforts, which typically involve identifying and normalizing genes being studied in an article. For instance, Mouse Genome Informatics (MGI) recently reported their search and evaluation of potential automatic tools for accelerating this gene finding process (4). Specifically, this year's GN task is to have participating systems return a list of gene database (Entrez Gene in this case) identifiers for a given article. There are two differences from past BioCreative GN challenges:

- Instead of using abstracts, full-length articles are provided.
- Instead of being species-specific, no species information is provided.

Both changes make this year's challenge event closer to the real literature curation task in MODs where humans are given full text articles without prior knowledge of organism information in the article.

Two additional new aspects of this year's GN task are the proposed evaluation metrics and the use of an EM algorithm for inferring ground truth based on team submissions. As many more genes are found in full text than in abstracts, returning genes by predicted confidence is preferred to a random order, as the former is more desirable in applications. Metrics used in past GN tasks such as Precision, Recall, and F-measure do not take ranking into consideration. Thus, we propose to use a new measure called

Threshold Average Precision (TAP-k), which is specifically designed for the measurement of retrieval efficacy in bioinformatics (13).

Finally, unlike in previous GN tasks where abstracts in the test set were completely hand annotated, the cost of manual curation on full text prevented us from obtaining human annotations for all 507 articles in the test set. Thus we resort to using team submissions for inferring ground truth. That is, given a labeling task and M independent labeling sources, it is possible to use these multiple sources to make estimates of the true labels which are generally more accurate than the labels from any single source alone. Perhaps the simplest approach to this is to use majority voting (5-7). On the other hand a number of methods have been developed using latent variables to represent in some way the quality of the labeling sources and based on the EM algorithm (8-12). There is evidence that such an approach can perform better than majority voting (8,11). We have chosen the most direct and transparent of the EM approaches (11) to apply to the GN task where we have multiple submissions as the multiple labeling sources. As far as we are aware this is the first attempt to base an evaluation of the performance of multiple computer algorithms on an EM algorithm for multiple independent data sources.

Methods

Data Preparation

For the purpose of obtaining full text articles in uniform formats and using them as a source for text analytics, all the articles selected for this task are published either by BioMed Central (BMC) or by Public Library of Science (PLoS), two PubMed Central (PMC) participating Open Access publishers. As a result, the text of each article was readily made available in both high-quality XML and PDF from PMC.

Participants were given a collection of training data to work with so that they could adjust their systems to optimal performance. The training set includes two sets of annotated full-length articles:

- 32 fully annotated articles by a group of invited professional MOD curators and by a group of bioinformaticians from the NCBI. Both groups were trained with detailed annotation guidelines (available as Appendix A) and a small number of example articles before producing gold-standard annotations. For each article in this set, a list of Entrez Gene ids is provided.
- A large number (500) of partially annotated articles. That is, not all genes that are mentioned in an article are annotated, but only the most important ones that within the scope of curation are annotated by human indexers at the National Library of Medicine (NLM). It is noted that most of the annotated genes are taken from the abstracts, though this is not 100%. This does not necessarily mean that the remainder of the text is useless. Presumably the full text can help to decide which genes are most important in the paper and determine the species to improve the prediction of the gene identifier.

For evaluating participating systems, we prepared a set of 507 articles as the test set. These articles were recently published and did not yet have any curated gene annotations. Due to the cost of manual curation, the same groups of curators were asked to produce human annotations only for a subset of 50 articles selected by the algorithm described below.

EM algorithm

In this scheme we assume there are M labeling sources and associate with the ith labeling source two numbers, the sensitivity as_i and the specificity bs_i. For the GN task we consider all the gene ids returned by the M sources as objects to be labeled.

Any given source produces a label for any such gene id which is the label "true" if the source returned that gene id or "false" if the source did not return that gene id. Then the sensitivity as_i is the probability that the ith source labels a correct gene id as true and the specificity bs_i is the probability that it labels an incorrect gene id as false. Assume there are N gene ids which require labeling. Then the model assumes a probability distribution $\{p_j\}_{j=1}^{N}$ where p_j is the probability that the jth gene id is correct. To begin the algorithm we initialize each p_j to be equal to the fraction of the M labels that are true for that gene id. The maximization step redefines the $\{as_i, bs_i\}_{i=1}^{M}$ in terms of the current $\{p_j\}_{j=1}^{N}$ by

$$\begin{aligned} as_i &= \left(1+\sum_{j=1}^{N}\delta_{ij}p_j\right)/\left(2+\sum_{j=1}^{N}p_j\right) \\ bs_i &= \left(1+\sum_{j=1}^{N}(1-\delta_{ij})(1-p_j)\right)/\left(2+\sum_{j=1}^{N}(1-p_j)\right) \end{aligned} \tag{0.1}$$

where we have used typical Laplace smoothing and define δ_{ij} to be 1 if the ith source labels the jth gene id as true and 0 otherwise. The p_js are defined for the subsequent expectation step by

$$p_j = \frac{pr_j\prod_{i=1}^{M}as_i^{\delta_{ij}}\left(1-as_i\right)^{\left(1-\delta_{ij}\right)}}{\left(pr_j\prod_{i=1}^{M}as_i^{\delta_{ij}}\left(1-as_i\right)^{\left(1-\delta_{ij}\right)}+(1-pr_j)\prod_{i=1}^{M}bs_i^{\left(1-\delta_{ij}\right)}\left(1-bs_i\right)^{\delta_{ij}}\right)} \tag{0.2}$$

by Bayes' theorem where for each j, pr_j is the prior for p_j. We initially took pr_j uniformly to be 0.5 and applied the algorithm to choose the 50 documents for hand labelling. Once we knew the correct annotations for the 50 document gold standard set we observed that only about 1% of gene ids returned by systems were correct. We subsequently have taken pr_j equal to 0.01 for all j in applying the algorithm to determine ground truth.

As mentioned above, our first use of this model was to find 50 documents among the 507 test documents which had the most variability in their labeling by different sources. For this purpose one submission from each team involved in the GN task was randomly selected and these submission were the 14 sources for application of the algorithm. When the algorithm was run to convergence we computed the entropy for the jth gene id by the formula

$$H_j = -p_j \log p_j - (1-p_j)\log(1-p_j) \tag{0.3}$$

Each document was scored by the sum of the entropies for all the gene ids coming from that document. Thus a document score is a function of how many gene ids are reported for that document and how variably the gene ids are reported by the different sources. This sampling, running the model and scoring the documents, was repeated 100 times and the top 50 documents varied only a small amount from run to run. We chose the 50 documents with the highest average scores over the 100 trials for hand annotation to provide the *gold standard* evaluation.

The second use of the model was to apply it to the best submission from each team. The choice of the best submission itself is based on the gold standard, but we made no further use of the gold standard. From the converged model using these sources we obtained a set of probabilities $\{p_j\}_{j=1}^{N}$ and we accepted as correct all those gene ids for which $p_j \geq 0.5$ and considered all other gene ids to be incorrect. This labeling we refer to as the *silver standard*. We used it to evaluate all submissions on the whole set of 507 documents. A comparison of results as computed with the gold standard and the silver standard is given in Table 3.

Evaluation Metrics

We propose to use a new metric, Threshold Average Precision (TAP-k), for evaluating team performance. In short, TAP is Mean Average Precision (MAP) with a variable cutoff and terminal cutoff penalty. We refer interested readers to the original publication (13) and Appendix B for detailed description of the TAP-k metric. In our evaluation, we used three values of k: 5, 10 and 20.

Results

GN Annotation Data

As shown in Table 1, the average numbers (mean and median) of annotated genes per article in Set 1 are significantly lower than the ones in Set 2, while remaining relatively close to its counterparts in Set 3. This comparison suggests that the 50 selected articles are not representative of the articles in the training set. Instead, the entire test set seems akin to the training set in this respect.

Table 1: Statistics of annotated gene ids in the different data sets.

Set	Description	Min	Max	Mean	Median	St.dev.
1	Training Set (32 articles)	4	147	19	14	24
2	Test Set (50 articles – gold standard)	0	375	33	19	63
3	Test Set (507 articles – silver standard)	0	375	18	12	27

Table 2 shows that there are many different species involved in this year's GN task, which suggests that species identification and disambiguation may be critical in the process of finding the correct gene ids. We also show that the distributions of species among the genes in the three data sets look largely different. This indeed reflects the method of selecting the articles for training and evaluation: with some prior knowledge of a papers' species information, we were able to select the 32 articles as the training set to match the domain expertise of those invited professional MOD curators in order to obtain best possible human annotations. On the other hand, the

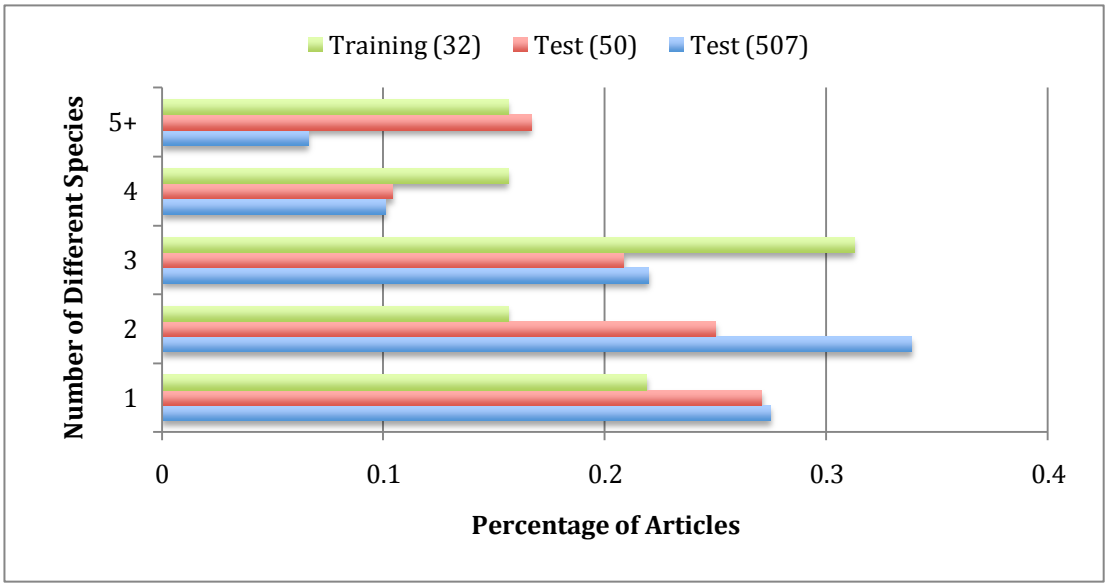
articles in the test set were selected rather randomly as none was annotated prior to the evaluation.

Table 2: Statistics of species distribution in the different data sets.

#	Training Set (32 articles)	Test Set (50 articles)	Test Set (507 articles)
1	S. cereviaiae (27%)	Enterobacter sp. 638 (23%)	H. Sapiens (42%)
2	H. sapiens (20%)	M. musculus (14%)	M. musulus (24%)
3	M. musculus (12%)	H. Sapiens (11%)	D. melanogaster (6%)
4	D. melanogaster (10%)	S. pneumoniae TIGR4 (9%)	S. cerevisiae S228c (6%)
5	D. rerio (7%)	S. scrofa (5%)	Enterobacter sp. 638 (4%)
6	A. thaliana (5%)	M. oryzae 70-15 (4%)	R. norvegicus (4%)
7	C. elegans (3%)	D. melanogaster (4%)	A. thaliana (2%)
8	X. laevis (3%)	R. norvegicus (3%)	C. elegans (2%)
9	R. norvegicus (2%)	S. cerevisiae S228c(2%)	S. pneumoniae TIGR4 (2%)
10	G. gallus (2%)	E. histolytica HM-1 (2%)	S. scrofa (1%)
11+	Other 18 species (9%)	Other 65 species (23%)	Other 91 species (7%)

In addition to recognizing various species in free text, participating systems also needed to properly link them to the corresponding gene mentions in the articles. As shown in Figure 1 most articles (over 70%) in our data sets contain more than one species mention. In fact, it is not uncommon to see 5 or more species in an article. In cases where more than one species is found in an article, it can be challenging for systems to associate a gene mention with its correct species.

Figure 1: Percentage of articles annotated with different numbers of species in various data sets. Training (32) refers to the human annotations on the 32 articles in the training set. Test (50) and Test (507) refer to the gold standard and silver standard annotations on the 50 and 507 articles in the test set, respectively.



Team Results

Each team was allowed to submit up to 3 runs. Overall, we received a total of 37 runs from 14 teams. One team withdrew their late submission (one run) before the results were returned to the teams. Thus, per their request we do not report their system performance in the tables below. Nevertheless we included their withdrawn run when selecting 50 articles and computing the silver standard by our EM algorithm, as we believe more team submission data are preferable in this case.

We assessed each submitted run by comparing it to the gold and silver standard, respectively, and report their corresponding TAP scores (k = 5, 10, and 20) in Table 3. As highlighted in the table, the two runs from team 83 (T83_R1 and T83_R3) achieved highest TAP scores in almost all cases except when evaluated on the silver standard with k = 20 where the third run from Team 98 (T98_R3) was the best. However, we did not find a statistically significant difference between the results of the two teams (T83 and T98) when comparing their respective best runs (with different values of k) based on the Wilcoxon signed rank test.

Table 3: Team evaluation results on the 50 and 507 articles using gold and silver standard annotations, respectively. Results are sorted by team numbers.

Team_Runs	Using gold standard (50 selected articles)			Using silver standard (All 507 articles)		
	TAP (K=5)	TAP K=10	TAP (K=20)	TAP (K = 5)	TAP (K = 10)	TAP (K = 20)
T63_R1	0.0337	0.0484	0.0718	0.1567	0.1939	0.1954
T63_R2	0.0296	0.0454	0.0638	0.1368	0.1855	0.1942
T65_R1	0.0628	0.0958	0.1017	0.1487	0.1754	0.1938
T65_R2	0.0891	0.1073	0.1156	0.1533	0.1817	0.2024
T68_R1	0.1568	0.1817	0.1987	0.3398	0.3551	0.3516
T68_R2	0.1255	0.1431	0.1740	0.3257	0.3410	0.3375
T70_R1	0.0566	0.0566	0.0566	0.1146	0.1146	0.1146
T70_R2	0.0622	0.0622	0.0622	0.1243	0.1243	0.1243
T70_R3	0.0718	0.0718	0.0718	0.1512	0.1512	0.1512
T74_R1	0.2099	0.2447	0.2447	0.4518	0.4518	0.4518
T74_R2	0.2045	0.2417	0.2417	0.4514	0.4514	0.4514
T74_R3	0.2061	0.2432	0.2432	0.4555	0.4555	0.4555
T78_R1	0.0577	0.0726	0.1106	0.1245	0.1527	0.1877
T78_R2	0.0829	0.1161	0.1662	0.2495	0.2655	0.2655
T78_R3	0.0830	0.1091	0.1387	0.2219	0.2645	0.2762
T80_R1	0.1072	0.1556	0.1622	0.3983	0.3983	0.3983
T80_R2	0.0372	0.0507	0.0578	0.2165	0.2165	0.2165
T80_R3	0.0324	0.0432	0.0516	0.2224	0.2288	0.2288
T83_R1	0.3184	**0.3469**	**0.3466**	**0.4581**	**0.4581**	0.4581
T83_R2	0.3147	0.3366	0.3366	0.4293	0.4293	0.4293
T83_R3	**0.3228**	0.3445	0.3445	0.4303	0.4303	0.4303
T89_R1	0.1197	0.1197	0.1351	0.2681	0.2989	0.2989
T89_R2	0.1351	0.1521	0.1620	0.2624	0.2950	0.2950
T89_R3	0.1275	0.1522	0.1522	0.2873	0.2873	0.2873
T93_R1	0.1599	0.1842	0.2010	0.3916	0.3916	0.3916
T93_R2	0.1517	0.1804	0.2000	0.3602	0.3720	0.3720
T93_R3	0.1611	0.1856	0.2032	0.3946	0.3946	0.3946
T97_R1	0.0709	0.092	0.1001	0.1369	0.1620	0.1859
T97_R2	0.0630	0.0849	0.0945	0.1304	0.1563	0.1770
T97_R3	0.0709	0.092	0.1001	0.1369	0.1620	0.1859
T98_R1	0.2805	0.2971	0.3064	0.3720	0.3802	0.3779
T98_R2	0.2850	0.3033	0.3044	0.3682	0.3775	0.3767
T98_R3	0.2973	0.3125	0.3248	0.4086	0.4511	**0.4648**
T101_R1	0.1849	0.2235	0.2331	0.4128	0.4128	0.4128
T101_R2	0.1649	0.2102	0.2365	0.4097	0.4224	0.4224
T101_R3	0.1773	0.2096	0.2374	0.4351	0.4351	0.4351

To assess the quality of the silver standard, we show in Table 4 the results of team submissions against the silver standard on the 50 selected articles. Although the two best runs from Team 83 in Table 3 are still among the ones with the highest TAP scores, they no longer are the best runs. Instead, the top positions are replaced by T74_R3 (for k=5) and T98_R3 (for k=10 and 20), respectively.

Table 4: Team evaluation results on the 50 articles using the sliver standard annotations. Results are sorted by team numbers.

Team_Run	TAP (K=5)	TAP (K=10)	TAP (K=20)
T63_R1	0.0515	0.1045	0.142
T63_R2	0.0455	0.0978	0.1335
T65_R1	0.0996	0.1259	0.1473
T65_R2	0.109	0.1317	0.1522
T68_R1	0.2238	0.2719	0.3152
T68_R2	0.2098	0.2917	0.2917
T70_R1	0.053	0.053	0.053
T70_R2	0.0566	0.0566	0.0566
T70_R3	0.096	0.096	0.096
T74_R1	0.3677	0.3677	0.3677
T74_R2	0.3713	0.3713	0.3713
T74_R3	**0.3747**	0.3747	0.3747
T78_R1	0.0589	0.0793	0.1139
T78_R2	0.1048	0.1548	0.2114
T78_R3	0.0972	0.1394	0.1949
T80_R1	0.2464	0.2719	0.2719
T80_R2	0.0663	0.1107	0.1177
T80_R3	0.0749	0.1231	0.1291
T83_R1	0.3498	0.3531	0.3531
T83_R2	0.3222	0.3222	0.3222
T83_R3	0.3313	0.3313	0.3313
T89_R1	0.1714	0.217	0.217
T89_R2	0.2141	0.2581	0.2949
T89_R3	0.2054	0.2054	0.2054
T93_R1	0.2518	0.2979	0.2979
T93_R2	0.2011	0.2514	0.2854
T93_R3	0.2487	0.293	0.293
T97_R1	0.1066	0.1307	0.149
T97_R2	0.09	0.1126	0.1323
T97_R3	0.1066	0.1307	0.149
T98_R1	0.3218	0.3388	0.3494
T98_R2	0.3217	0.3391	0.3496
T98_R3	0.3576	**0.3953**	**0.4499**
T101_R1	0.3504	0.374	0.374
T101_R2	0.3068	0.379	0.3976
T101_R3	0.3077	0.3942	0.3942

Discussion

Team Results and Quality of Silver Standard

Although we are unable to directly compare this year's GN team results against the ones in previous GN challenges due to different evaluation metrics, results in Table 3

led us to believe that this year's GN task is more challenging, potentially due to the complexity of full text processing and species identification (14,15).

Using the silver standard allowed us to assess team submissions on the entire set of test articles without having human annotations for all articles. As can be seen from results in Tables 3 and 4, TAP scores are consistently higher when evaluated on the silver standard compared to the gold standard. Furthermore, individual team rankings may be affected. For instance, as mentioned earlier the best performing run was T83_R3 using gold standard but T74_R3 using silver standard. Nevertheless, it is evident that relative rankings tend to be largely preserved in this comparison. For instance, teams 83, 74, 98 and 101 consistently remain as the top tier group in all evaluations. This provides some justification for the silver standard and suggests that this approach to evaluation has some merit.

As just noted, TAP scores in Table 3 show that overall team performance is lower on the 50 articles than on the entire set of 507 articles. The reasons for this are two fold. First, the 50 articles are the most difficult ones for gene normalization (as shown by comparing the silver results for the 50 and the 507) and this supports our rationale for their choice. Second, by comparing the gold and silver results for the 50 in Tables 3 and 4, we can see that team results are always higher when evaluated using the silver standard. Taken together, this suggests that the true TAP scores on the entire test set should be slightly lower than what is currently reported using the silver standard in Table 3.

Team Methods

Each team was required to submit a system description before receiving the gold standard annotations on the 50 articles and their scores. Based on reading those

submitted descriptions, we found the general framework for the gene normalization task comprises the following major steps:

1) Identifying gene mentions
2) Identifying species information and linking such information to gene mentions
3) Retrieving a list of candidate gene ids for a given gene mention
4) Selecting gene ids through disambiguation.

Conclusions

We have successfully organized a community-wide challenge event for the gene normalization task. There were a total of 37 submissions by 14 different teams from Asia, Europe, and North America. The highest TAP-k scores obtained on the gold-standard annotations of the 50 test articles are 0.3248 (k=5), 0.3469 (k=10), and 0.3466 (k=20), respectively. In addition, TAP-k scores of 0.4581 (k=5, 10) and 0.4684 (k=20) are observed when using the silver standard of the 507 test articles. In comparison with past BioCreative GN tasks, this year's task bears more resemblance to real-world tasks in which curators are given full text without knowing species information. As a consequence, this year's task has proved more difficult than the ones in the past, which is evident from the overall lower team performance.

Finally, we believe the TAP-k metric and EM algorithm proved to be adequate for evaluating retrieval efficacy and for inferring ground truth based on team submissions. In particular, the proposed pooling method allowed us to effectively detect good team performance without having to relying on human annotations.

Future work should include conducting a more detailed analysis of various techniques and tools used by different participating teams, as this may provide valuable direction for future research on the GN problem. Also, we plan to combine results from different teams as an ensemble system to test maximal aggregate performance, as in

various previous studies (1,16,17). Finally, we would like to investigate how systems developed for the GN task may be used in real-world applications.

Additional material

Additional file 1: GN annotation guidelines

Additional file 2: Introduction to TAP-k

Acknowledgements

This research is supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors would like to thank all the annotators who produced the gold-standard annotations.

References

1. Morgan, A.A., Lu, Z., Wang, X., *et al.* (2008) Overview of BioCreative II gene normalization. *Genome Biol*, 9 Suppl 2, S3.
2. Hirschman, L., Colosimo, M., Morgan, A., Yeh, A. (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, 6 Suppl 1, S11.
3. Colosimo, M.E., Morgan, A.A., Yeh, A.S., Colombe, J.B., Hirschman, L. (2005) Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*, 6 Suppl 1, S12.
4. Dowell, K.G., McAndrews-Hill, M.S., Hill, D.P., Drabkin, H.J., Blake, J.A. (2009) Integrating text mining into the MGI biocuration workflow. *Database (Oxford)*, 2009, bap019.
5. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y. (2008) Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii.
6. Sheng, V.S., Provost, F., Ipeirotis, P.G. (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Las Vegas, Nevada, USA.
7. Donmez, P., Carbonell, J.G., Schneider, J. (2009) Efficiently learning the accuracy of labeling sources for selective sampling. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Paris, France.
8. Whitechill, J., Ruvolo, P., Wu, T., Bergsma, J., Movellan, J. (2009) Whose vote should count more: optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems*, 2035-3043.
9. Welinder, P., Perona, P. (2010) Online crowdsourcing: rating annotators and obtaining cost-effective labels. *Workshop on Advancing Computer Vision with Humans in the Loop at CVPR'10*.

10. Smyth, P., Fayyad, U., Burl, M., Perona, P., Baldi, P. (1995) Inferring ground truth from subjective labelling of venus images. *Advances in Neural Information Processing Systems*, 7.
11. Raykar, V.C., Yu, S., Zhao, L.H., *et al.* (2010) Learning From Crowds. *Journal of Machine Learning Research*, 11, 1297-1322.
12. Dawid, A.P., Skene, A.M. (1979) Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28, 20-28.
13. Carroll, H.D., Kann, M.G., Sheetlin, S.L., Spouge, J.L. (2010) Threshold Average Precision (TAP-k): a measure of retrieval designed for bioinformatics. *Bioinformatics*, 26, 1708-1713.
14. Kappeler, T., Kaljurand, K., Rinaldi, F. (2009) TX task: automatic detection of focus organisms in biomedical publications. *Proceedings of the Workshop on BioNLP*. Association for Computational Linguistics, Boulder, Colorado.
15. Wang, X., Tsujii, J., Ananiadou, S. (2010) Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics*, 26, 661-7.
16. Leitner, F., Mardis, S.A., Krallinger, M., Cesareni, G., Hirschman, L.A., Valencia, A. (2010) An Overview of BioCreative II.5. *IEEE/ACM Trans Comput Biol Bioinform*, 7, 385-399.
17. Smith, L., Tanabe, L.K., Ando, R.J., *et al.* (2008) Overview of BioCreative II gene mention recognition. *Genome Biol*, 9 Suppl 2, S2.

BioCreative GN Task 2010

Gene/Protein Annotation Guidelines

What to annotate and normalize:

1. Find gene/protein mentions in the full-length article including figure and table legends and map them to unique Entrez Gene identifiers (http://www.ncbi.nlm.nih.gov/gene/).
2. Entrez Gene Ids are required. (UniProt Ids or Model Organism Database Ids are optional).
3. Annotate all genes mentioned in the article including those genes mentioned in passing or only mentioned once in the article. However, there is no need to rank or group genes for this assignment.
4. When there is no explicit mention of a gene's organism of origin in surrounding text, try to use the article context to help determine its species. Annotate the gene only when the species information can be determined. Some helpful clues for determining species include details in the methods/materials section such as cell lines, organism-specific gene nomenclature conventions, etc.
5. You may also use your domain knowledge for determining which organism a gene belongs to when no explicit species information is given in the text. If there is absolutely no clue about the species, or in situations where the species information is ambiguous (e.g. the authors use one gene as a representative of its homologs), do not annotate the gene.
6. When cell lines from different species are used to study a gene, determine and use the gene's *species of origin* instead of a cell lines' *species of origin* for annotation.

What NOT to annotate:

1. Do not annotate references sections. But this section may be useful for species identification. However, do not go beyond reading reference titles. That is, don't read the referenced articles.
2. Do not use or annotate supplementary material or supporting information.
3. Annotate target proteins but do not annotate antibodies/reagents that are used to study target proteins.
4. Do not annotate the Methods/Materials section for genes/proteins. But this section may be useful for species identification. (Our reasoning is that the Methods/Materials section often contains information about reagents or antibodies that are themselves proteins but are not *curatable* objects; if *curatable* genes/proteins are mentioned in such a section, then they will almost certainly be mentioned elsewhere in the article).
5. Do not annotate genes where no unique ids can be identified in Entrez Gene. For example, if you find a gene mention "x-tsk" in a paper and subsequently search it in Entrez Gene, you may be presented with two separate Entrez gene records (x-tsk-b1 & x-tsk-b2). In this case, if you can't tell which specific gene is used in the paper based on your domain knowledge, do not annotate this gene.
6. Do not annotate a protein complex (e.g. TFTC complex). But if its members are explicitly given (NFKB-IKB complex) they should be annotated.
7. Do not annotate a protein family (e.g. cytokines; ring-h2 finger proteins) because no unique Entrez Gene id can be assigned to it.
8. Do not annotate a gene/protein with only non-species taxonomic information (e.g. mammalian p53) for the same reason above.

What is *TAP-k*?

Here we refer to the measure defined by Carroll, H. D., Kann, M. G., Sheetlin, S. L., and Spouge, J. L., Threshold Average Precision (TAP-k): A Measure of Retrieval Designed for Bioinformatics, *Bioinformatics Advanced Access published on May 26, 2010*.

The Threshold Average Precision (*TAP-k*) is *MAP* with a variable cutoff and terminal cutoff penalty.

For a single query the average precision (*AP*) is computed by summing the precision at each rank that contains a true positive item and then dividing this sum by the number of positives for that query. If the retrieval system assigns to each retrieved item a score and the retrieved items are ranked in decreasing order of score, then it may be useful to cut off the retrieval at some fixed score level x. We can compute the average precision with cutoff x (APC_x). This is the sum of the precision at each rank with a true positive item and a score $>=x$, divided by the total number of positives for the query. Finally, suppose that $y>x$ and further suppose there are no true positive items in the sum for APC_x that are below y. Then $APC_y=APC_x$. But clearly it would make more sense to choose the cutoff y than the cutoff x. To distinguish between these two cases we define the average precision with cutoff x and terminal penalty ($APCP_x$). Let P_x be the precision at the last rank with score $>= x$ and let P be the total number of positives. Then define

$$APCP_x = \frac{TP * APC_x + 1 * P_x}{TP + 1}. \qquad (1.1)$$

$APCP_x$ is just the weighted average of APC_x and P_x with most of the weight applied to APC_x, but P_x supplying the terminal penalty. In our hypothetical case P_y will be greater than P_x so that $APCP_y$ is also greater than $APCP_x$ and the score rewards the better choice of cutoff or equally penalizes the poorer choice. Whereas *MAP* is the average of *AP* over all the queries, *TAP-k* is the average of $APCP_x$ over all the queries where x is chosen as the largest score that produces a median of k false positive retrievals over all the queries. The median is used here instead of the mean because it is more robust against noise and outliers.

There are some practical considerations when applying *TAP-k*. First, retrieval systems must produce scores commensurate with their rankings and these scores must be interpretable across different queries. Since most systems generate their retrieval by scoring this should not make the task any more difficult than usual. On the other hand some kind of score normalization may be necessary for some systems, depending on how the scores are constructed. An ideal score would be a probability estimate that the retrieved item is a true positive, but a score need not be a probability estimate for good performance. The score that is reported simply has to have the same implications for relevance of the item regardless of the query, for the best performance. Another important issue is the length of the retrieved lists returned by a system. If many of the retrieved lists are too short to have k false positives appear, then no cutoff score may produce a

median number of k false positive retrievals for the set of queries. In that case we will take the cutoff score x to be the lowest score over all the retrieval lists for all the queries.

Example 1. Data for five queries, Q1-Q5 are presented in the table. The numbers in parentheses following the query numbers are the number of correct or relevant items for each query. This data was generated randomly based on the scores. Each score is the probability that the corresponding retrieved item would be relevant (relevance is shown by a 1 in the rel column for each query). The scores themselves are parts of power series which are convenient for generating realistic scores. Retrieval is cut off at 15 items for each query to keep the data easily manageable and as a consequence not all relevant items are necessarily retrieved.

	Q1 (5)		Q2 (5)		Q3 (5)		Q4 (3)		Q5 (5)	
	rel	score	rel	score	rel	score	rel	score	rel	score
1	1	0.900	0	0.500	0	0.500	0	0.2	1	0.980
2	1	0.738	0	0.475	1	0.475	0	0.187	0	0.788
3	0	0.605	1	0.451	0	0.451	0	0.174	0	0.633
4	1	0.496	0	0.429	0	0.429	0	0.163	1	0.509
5	1	0.407	1	0.407	0	0.407	0	0.152	1	0.409
6	0	0.334	0	0.387	0	0.387	0	0.142	0	0.329
7	0	0.274	0	0.367	0	0.367	0	0.132	0	0.265
8	0	0.224	0	0.349	1	0.349	0	0.123	0	0.213
9	1	0.184	0	0.332	0	0.332	0	0.115	0	0.171
10	0	0.151	1	0.315	1	0.315	0	0.107	1	0.138
11	0	0.124	0	0.299	0	0.299	0	0.100	0	0.111
12	0	0.101	0	0.284	0	0.284	0	0.094	0	0.089
13	0	0.083	0	0.270	0	0.270	0	0.087	0	0.071
14	0	0.068	0	0.257	0	0.257	0	0.082	0	0.057
15	0	0.056	0	0.244	1	0.244	0	0.076	0	0.046

Here the score cutoff for *TAP*-5 is 0.213 and the values of $APCP_5$ are 0.675, 0.206, 0.264, 0, 0.413 and the average of these numbers, *TAP*-5, is 0.312. The blue background shows what parts of the retrieval were included in the scoring (likewise for subsequent examples).

Example 2. Example 1 output, but the system has limited its retrieval to the top 4 ranks for each query.

	Q1 (5)		Q2 (5)		Q3 (5)		Q4 (3)		Q5 (5)	
	rel	score	rel	score	rel	score	rel	score	rel	score
1	1	0.900	0	0.500	0	0.500	0	0.2	1	0.980
2	1	0.738	0	0.475	1	0.475	0	0.187	0	0.788
3	0	0.605	1	0.451	0	0.451	0	0.174	0	0.633
4	1	0.496	0	0.429	0	0.429	0	0.163	1	0.509
5										
6										
7										
8										
9										
10										

11										
12										
13										
14										
15										

Here the cutoff score is 0.163 (the lowest score possible) and the $APCP_5$ values are 0.583, 0.097, 0.125, 0, 0.333 and the average, *TAP*-5, of these numbers is 0.228. Here the *TAP*-5 is lower than for example 1 because the system cut the retrieval off prematurely and this decreased the recall and thus the *TAP* -5 score.

Example 3. Example 1 output again, but scores changed so they only reflect the rank and not the quality of the retrieved material.

	Q1 (5)		Q2 (5)		Q3 (5)		Q4 (3)		Q5 (5)	
	rel	score	rel	score	rel	score	rel	score	rel	score
1	1	0.9	0	0.9	0	0.9	0	0.9	1	0.9
2	1	0.85	0	0.85	1	0.85	0	0.85	0	0.85
3	0	0.8	1	0.8	0	0.8	0	0.8	0	0.8
4	1	0.75	0	0.75	0	0.75	0	0.75	1	0.75
5	1	0.7	1	0.7	0	0.7	0	0.7	1	0.7
6	0	0.65	0	0.65	0	0.65	0	0.65	0	0.65
7	0	0.6	0	0.6	0	0.6	0	0.6	0	0.6
8	0	0.55	0	0.55	1	0.55	0	0.55	0	0.55
9	1	0.5	0	0.5	0	0.5	0	0.5	0	0.5
10	0	0.45	1	0.45	1	0.45	0	0.45	1	0.45
11	0	0.4	0	0.4	0	0.4	0	0.4	0	0.4
12	0	0.35	0	0.35	0	0.35	0	0.35	0	0.35
13	0	0.3	0	0.3	0	0.3	0	0.3	0	0.3
14	0	0.25	0	0.25	0	0.25	0	0.25	0	0.25
15	0	0.2	0	0.2	1	0.2	0	0.2	0	0.2

Here the scores no longer reflect quality and thus they do not give an accurate idea of where to cut off retrieval to obtain maximal efficiency. As a result there is a drop in *TAP*-5 as compared with example 1. The cutoff score is 0.6 and the $APCP_5$ values are 0.687, 0.170, 0.107, 0, 0.421 and the average, *TAP*-5, is 0.277.